MULTILEVEL MONTE CARLO FOR SMOOTHING VIA TRANSPORT METHODS

JEREMIE HOUSSINEAU*, AJAY JASRA†, AND SUMEETPAL S. SINGH‡

Abstract. In this article we consider recursive approximations of the smoothing distribution associated to partially observed stochastic differential equations (SDEs), which are observed discretely in time. Such models appear in a wide variety of applications including econometrics, finance and engineering. This problem is notoriously challenging, as the smoother is not available analytically and hence require numerical approximation. This usually consists by applying a time-discretization to the SDE, for instance the Euler method, and then applying a numerical (e.g. Monte Carlo) method to approximate the smoother. This has lead to a vast literature on methodology for solving such problems, perhaps the most popular of which is based upon the particle filter (PF) e.g. [9]. In the context of filtering for this class of problems, it is well-known that the particle filter can be improved upon in terms of cost to achieve a given mean squared error (MSE) for estimates. This in the sense that the computational effort can be reduced to achieve this target MSE, by using multilevel (ML) methods [12, 13, 18], via the multilevel particle filter (MLPF) [16, 20, 21]. For instance, to obtain a MSE of $\mathcal{O}(\epsilon^2)$ for some $\epsilon > 0$ when approximating filtering distributions associated with Euler-discretized diffusions with constant diffusion coefficients, the cost of the PF is $\mathcal{O}(\epsilon^{-3})$ while the cost of the MLPF is $\mathcal{O}(\epsilon^{-2}\log(\epsilon)^2)$. In this article we consider a new approach to replace the particle filter, using transport methods in [27]. In the context of filtering, one expects that the proposed method improves upon the MLPF by yielding, under assumptions, a MSE of $\mathcal{O}(\epsilon^2)$ for a cost of $\mathcal{O}(\epsilon^{-2})$. This is established theoretically in an "ideal" example and numerically in numerous examples.

Key words. Transport map, Stochastic differential equation, Multilevel Monte Carlo

AMS subject classifications. 62M05, 60J60

1. Introduction.

The smoothing problem often refers to the scenario where one has an unobserved Markov chain (or signal) in discrete or continuous time and one is interested in inferring the hidden process on the basis of observations, which depend upon the hidden chain. The case we consider is where the hidden process follows a SDE and the observations are regularly recorded at discrete times; given the signal at a time t the observation is assumed to be conditionally independent of all other random variables. The process of filtering is to infer some functional of the hidden state at time t given all the observations at time t and the smoothing problem to infer some functional of potentially all the states at the discrete observation times again given all the observations. It is often of interest to do this recursively in time. This modelling context is relevant for many real applications in econometrics, finance and engineering; see e.g. [4] and the references therein.

The smoothing problem is notoriously challenging. Supposing one has access to the exact transition of the SDE, then unless the observation density is Gaussian and depends linearly on the hidden state and the transition density is also Gaussian depending linearly on the previous state, the filter and smoother are not analytically tractable (unless the state-space of the position of the diffusion at any given time is finite and of small cardinality); see [5]. However, it is seldom the case that even the transition density (or some unbiased approximation of it, e.g. [11] and the references therein) is available; this is assumed throughout the article. Thus typically, one time-discretizes the diffusion process and then one seeks to perform filtering and

*DSAP, National University of Singapore. Email: stahje@nus.edu.sg

†DSAP, National University of Singapore. Email: staja@nus.edu.sg

‡Department of Engineering, University of Cambridge and The Alan Turing Institute. Email: sss40@cam.ac.uk

smoothing from the time-discretized model. This latter task is still challenging as it is still analytically intractable. There is a vast literature on how to numerically approximate the filter/smoother (e.g. [7]) and perhaps the most popular of which is the particle filter. This is a method whose cost grows linearly with the time parameter and generates N samples in parallel. These samples are put through sampling and resampling operations. It is well-known that when estimating the filter, the error is uniform in time. For the smoother, the error often grows due to the so-called path degeneracy problem and indeed, there are many smoothing problems for which it is not appropriate; see [23] for some review and discussion. In the context of the problem in this article, when only considering the filter, ignoring the time parameter and under assumptions, to obtain a MSE of $\mathcal{O}(\epsilon^2)$ for some $\epsilon > 0$ the cost of the PF is $\mathcal{O}(\epsilon^{-3})$. The MSE takes into account the exact filter (i.e. the one with no time discretization).

Multilevel Monte Carlo (MLMC) methods [12, 13, 14, 15, 18] are of interest in continuum systems which have to be discretized in one dimension, just as in this article (extensions to discretization in multiple dimensions have been proposed and studied in [6, 17]). We explain the idea informally as follows: let the time parameter be fixed and denote by p_t^L the filter associated to a (say Euler) discretization level $h_L > 0$, set $X_t \in \mathbb{R}^d$, $d \geq 1$ and for $\varphi : \mathbb{R}^d \to \mathbb{R}$ bounded denote by $p_t^L(\varphi)$ the expectation of φ with respect to (w.r.t.) the filter. Then the MLMC method is based upon the following approach. Consider $0 < h_L < h_{L-1} < \cdots < h_0 < +\infty$ a sequence of discretizations, where h_L is the most accurate (finest) discretization and h_0 the least (coarsest), the ML identity is

$$p_t^L(\varphi) = \sum_{l=0}^{L} (p_t^l - p_t^{l-1})(\varphi)$$

where p_t^{-1} is an arbitrary measure satisfying $p_t^{-1}(\varphi) = 0$ for every φ. The idea is then to sample N_0 independent samples from p_t^0 and then, independently for each $1 \leq l \leq L$ independently sample N_l coupled pairs from the pair (p_t^l, p_t^{l-1}). The MLMC estimator is then

$$\frac{1}{N_0} \sum_{i=1}^{N_0} \varphi(X_{t,i}^0) + \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} [\varphi(X_{t,i}^l) - \varphi(X_{t,i}^{l-})]$$

where $\{X_{t,i}^0\}_{i=1}^{N_0}$ are i.i.d. p_t^0 and $\{(X_{t,i}^l, X_{t,i}^{l-})\}_{i=1}^{N_l}$ are i.i.d. from a coupling of (p_t^l, p_t^{l-1}). To obtain a MSE of $\mathcal{O}(\epsilon^2)$ one sets L such that the squared bias is $\mathcal{O}(\epsilon^2)$ (the bias is known in the context of interest). If one has $\mathrm{Var}(\varphi(X_{t,1}^l) - \varphi(X_{t,1}^{l-})) = \mathcal{O}(h_l^\beta)$ for some $\beta > 0$ then one can try to minimize (w.r.t. $N_1, \ldots, N_L$) the cost $\sum_{l=1}^L N_l h_l^{-\zeta}$ ($\zeta = 1$ for an Euler discretization) subject to the variance $1/N_0 + \sum_{l=1}^L h_l^\beta / N_l$ being $\mathcal{O}(\epsilon^2)$. [12] finds a solution to this problem. The main issue in the context of smoothing, is that one (typically) does not know how to sample from the smoothers nor the couplings.

In [16, 20, 21] it is shown how to utilize the PF to leverage on the potential decrease in cost to obtain a given MSE. This has been termed the MLPF. The idea is to use couplings in the Euler dynamics and the resampling operation of a PF. This has been later refined in [26]. To our knowledge, the only theoretical work for the MLPF in [20], shows that to obtain a MSE of $\mathcal{O}(\epsilon^2)$ the cost in MLPF is $\mathcal{O}(\epsilon^{-2} \log(\epsilon)^2)$, for some specific (constant diffusion coefficient) models and under particular assumptions. This is known to be worse than the rates obtained in [12] in the case where there are no

observations. Here and throughout, the time parameter is omitted from the discussion on cost and error, despite the fact that these are important considerations in general.

The main idea in this article is to adopt an alternative method to the PF. The approach is to use transport methods [27]. Transport maps have been used for Bayesian inference [10, 19] and more specifically for parameter estimation in [24] based on a related multi-scale idea. The basic idea is to obtain a map such that the image of samples from an easy-to-sample distribution through this map has exactly the type which one desires. In [27] it is shown how to develop numerical approximations of maps, associated exactly to the distributions of interest in this article. These approximations often induce i.i.d. Monte Carlo approximations of expectations of interest, albeit with a numerical error associated to the approximation of the transport map. As mentioned in [22], it is simple to induce coupled pairs using the method of [27] and this is exactly what is done in this paper. The potential advantages of this method relative to the MLPF are then as follows:

(i) The ML rate lost by coupled resampling can be regained in the context of filtering.

(ii) The method can be used for approximating the expectation of some functionals w.r.t. the smoother, whereas the approach in [20, 21] is typically not useful for smoothing at large time-lags.

In this article we establish that (i) can hold in an ideal special case, where the model is linear and Gaussian and the transport map is exact. This result is reinforced by numerical examples which show that the result seems to hold more generally. The significance of (i) is that to obtain a MSE of $\mathcal{O}(\epsilon^2)$ the cost is $\mathcal{O}(\epsilon^{-2})$; this is better than the MLPF. Point (ii) relates to the afore-mentioned path degeneracy effect, which can mean PFs (and hence the MLPF) are not so useful in the context of large lag smoothing.

The structure of the article is as follows: Section 2 introduces the model and transport methodology. Section 3 presents the multilevel approach and the MLPF as well as the mechanisms underlying the computation of transport maps for a given level of discretization. The efficiency of the proposed approach is shown numerically on increasingly challenging scenarios in section 4.

2. Methodology for SDE smoothing.

In this section, the considered notations and assumptions for the smoothing of SDEs are presented, together with a brief overview of the transport methodology.

2.1. The SDE model.

Throughout the article, all random variables will be assumed to be on the same complete probability space $(\Omega, \Sigma, \mathbb{P})$ and will be denoted by upper-case letters, while their realisations will be in lower case. We consider a diffusion process $\boldsymbol{X} = \{X_t\}_{t\in[0,T]}$ on the space $\mathbb{R}^d$ of the form

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t + b(X_t)\mathrm{d}W_t, \qquad t \in [0,T], \tag{2.1}$$

where T is the final time, $\{W_t\}_{t\in[0,T]}$ is the Brownian motion on $\mathbb{R}^d$, $a(\cdot)$ is in the set $\mathcal{C}^2(\mathbb{R}^d, \mathbb{R}^d)$ of twice continuously differentiable mappings from $\mathbb{R}^d$ to itself and $b(\cdot)$ is in $\mathcal{C}^2(\mathbb{R}^d, \mathbb{M}_d(\mathbb{R}))$ with $\mathbb{M}_d(\mathbb{R})$ the space of square matrices of size d. The mapping b is assumed to be such that $b(x)b(x)^{\mathrm{t}}$ is positive definite for all $x \in \mathbb{R}^d$, with $\cdot^{\mathrm{t}}$ denoting the transposition. Moreover, the drift and diffusion coefficients are assumed to be globally Lipschitz, i.e. there exists $c > 0$ such that

$$|a(x) - a(x')| + |b(x) - b(x')| \leq c|x - x'|$$

for all $x, x' \in \mathbb{R}^d$. The initial distribution of the process $\boldsymbol{X}$, i.e. the distribution of X_0, is denoted p_0 (and might be equal to δ_{x_0} for some initial condition $x_0 \in \mathbb{R}^d$). It is assumed that the m^{th}-order moment of X_0 defined as $\mathbb{E}(|X_0|^m)$ is finite for any $m \geq 1$. Probability density functions will be considered with respect to the Lebesgue measure on $\mathbb{R}^d$ and both probability measures and their corresponding density functions will be referred to by the same notation.

The distribution of X_k, $k \in \{1, \dots, T\}$, given a realisation x_{k-1} of the state X_{k-1} is denoted $Q(x_{k-1}, \cdot)$. In addition to the fact that the expression of the Markov transition Q is unavailable in general, it is not usually possible to devise an unbiased estimator for it or even to sample from it. In the case where $d = 1$, one can obtain "skeletons" of exact paths using the algorithm of [3, 2], however, the extension of this approach to SDEs of higher dimensions might not be possible [1].

The diffusion process $\boldsymbol{X}$ is assumed to be observed in $\mathbb{R}^{d'}$, $d' \in \mathbb{N}$, at all the integer-valued times so that the final time T is also assumed to be an integer. These assumptions are made for the sake of notational simplicity and can be easily removed. For all $k \in \{0, \dots, T\}$, the observation Y_k is a random variable that is conditionally independent on the state X_t at times $t \neq k$ given X_k. The observation process can be expressed in general as

$$Y_k = g_k(X_k, V_k) \tag{2.2}$$

where g_k is a deterministic observation function and where $\{V_k\}_{k=0}^T$ is a collection of independent random variables. It is assumed without any real loss of generality that both g_k and the distribution of V_k do not depend on the time index k, the corresponding likelihood for a realisation y_k of Y_k is denoted $\ell(X_k, y_k)$.

2.2. Smoothing for SDEs. Throughout the article, joint states in $\mathbb{R}^{d(n+1)}$ for some $n \in \mathbb{N}_0$ will be denoted either by $x_{k:k+n} \doteq (x_k, x_{k+1}, \dots, x_{k+n})$ with $k \in \mathbb{N}$ or by x_S, with $S = \{s_0, s_1, \dots, s_n\}$ a finite subset of $[0, T]$ such that $s_i < s_j$ for all $0 \leq i < j \leq n$, defined as $x_S \doteq (x_{s_1}, x_{s_2}, \dots, x_{s_n})$. The smoothing distribution associated with the SDE (2.1) is defined formally as the joint law of the diffusion process $\boldsymbol{X}$ at all the integer times given realisations $y_0, \dots, y_T$ of the observation process (2.2), and can be expressed for any $x_{0:T} \in \mathbb{R}^{d(T+1)}$ as

$$\boldsymbol{p}(x_{0:T}) = \frac{\ell(x_0, y_0) p_0(x_0) \prod_{k=1}^T \left[Q(x_{k-1}, x_k) \ell(x_k, y_k)\right]}{\int \ell(x'_0, y_0) p_0(x'_0) \prod_{k=1}^T \left[Q(x'_{k-1}, x'_k) \ell(x'_k, y_k)\right] \mathrm{d}x'_{0:T}}.$$

The dependence of the smoothing distribution on the realisations $y_0, \dots, y_T$ of the observation process is omitted for the sake of notational simplicity. This is justified by the fact that these observations will be fixed in the remainder of the article so that the smoothing distribution $\boldsymbol{p}$ and its approximations will always be conditioned on the same given observations. The expression of $\boldsymbol{p}$ is a direct consequence of Bayes' theorem applied to the prior $p_0(x_0) \prod_{k=1}^T Q(x_{k-1}, x_k)$ describing the law of the unobserved (hidden) diffusion process together with the joint likelihood $\prod_{k=0}^T \ell(x_k, y_k)$ whose expression results from the conditional independence of the observations.

Using the same principle of implicit conditioning as with the smoothing distribution, the filtering distribution p_k at time k is defined as the law of X_k given the realisations $y_0, \dots, y_k$ and is expressed recursively as

$$p_k(x_k) = \frac{\ell(x_k, y_k) \int Q(x_{k-1}, x_k) p_{k-1}(x_{k-1}) \mathrm{d}x_{k-1}}{\int \ell(x'_k, y_k) Q(x'_{k-1}, x'_k) p_{k-1}(x'_{k-1}) \mathrm{d}x'_k \mathrm{d}x'_{k-1}}$$

for any $x_k \in \mathbb{R}^d$ and any $k \in \{1, \dots, T\}$. The marginal distribution of X_k induced by the smoothing distribution $\boldsymbol{p}$ corresponds to the filtering distribution p_k when $k = T$ only.

The objective in this article can now be formally expressed as follows: to compute the expectation $\boldsymbol{p}(\varphi) \doteq \int \varphi(x_{0:T})\boldsymbol{p}(x_{0:T})\mathrm{d}x_{0:T}$ of some bounded measurable function φ on $\mathbb{R}^{d(T+1)}$. Although the above formulation casts the considered problem into the standard Bayesian inference framework, the Markov transition Q is unavailable in general, so that expressing analytically the distributions $\boldsymbol{p}$ and p_k is not usually possible. The first step toward our objective is then to apply a time-discretization to the SDE (2.1), which, for the sake of simplicity, is illustrated with Euler's method for some discretization level $l \in \mathbb{N}_0$:

$$X_{t+h_l} = X_t + h_l a(X_t) + \sqrt{h_l} b(X_t) U_t, \tag{2.3}$$

for some time-step $h_l = 2^{-l}$ and for all $t \in \mathcal{T}_l \setminus \{T\}$ where $\mathcal{T}_l \doteq \{0, h_l, \dots, T\}$, with $\{U_t\}_{t \in \mathcal{T}_l \setminus \{T\}}$ a collection of independent Gaussian random variables with density $\phi(\cdot\,; 0, \boldsymbol{I}_d)$ where $\boldsymbol{I}_d$ is the identity matrix of size d. The choice of time step $h_l = 2^{-l}$ is made for the sake of convenience and is not necessary. The only requirement for both the MLPF and the multilevel transport is that the ratio h_{l-1}/h_l has to be an integer. The number of time steps from a given observation time up to and including the next observation time, that is in the interval $(k, k+1]$ for some $k \in \{0, \dots, T-1\}$, is $M_l = 2^l$. The numeral scheme (2.3) yields a Markov transition K^l between two successive discretization times defined as

$$K^l(x, \cdot) = \phi\big(\cdot\,; x + h_l a(x), h_l b(x) b(x)^{\mathrm{t}}\big)$$

for any $x \in \mathbb{R}^d$, which enables the approximation of Q by another Markov kernel Q^l defined as

$$Q^l(x, \cdot) = \underbrace{K^l \dots K^l}_{M_l \text{ times}}(x, \cdot),$$

where $KK'(x, \cdot) = \int K(x, x')K'(x', \cdot)\mathrm{d}x'$ for any transition kernels K, K'. The smoothing distribution $\boldsymbol{p}^l$ induced by (2.3), which approximates $\boldsymbol{p}$, is expressed on $\mathbb{R}^{d(M_l T+1)}$ instead of $\mathbb{R}^{d(T+1)}$ and is characterised by

$$\boldsymbol{p}^l(x_{\mathcal{T}_l}) \propto p_0(x_0) \prod_{t \in \mathcal{T}_l \setminus \{T\}} K^l\big(x_t, x_{t+h_l}\big) \prod_{k=0}^{T} \ell(x_k, y_k)$$

for any $x_{\mathcal{T}_l} \in \mathbb{R}^{d(M_l T+1)}$. Marginalising w.r.t. all x_t such that $t \notin \mathbb{N}_0$ gives a distribution on $\mathbb{R}^{d(T+1)}$ which depends on the same time steps as $\boldsymbol{p}$. It is understood that the error in the approximation of Q and $\boldsymbol{p}$ by Q^l and $\boldsymbol{p}^l$ decreases when l increases and tend to 0 as l tends to infinity. The measure $\boldsymbol{p}^l(\varphi)$ of the function φ is understood as the measure of the canonical extension $\bar{\varphi}$ of φ from $\mathbb{R}^{d(T+1)}$ to $\mathbb{R}^{d(M_l T+1)}$ defined as

$$\bar{\varphi}(x_t) = \begin{cases} \varphi(x_t) & \text{if } t \in \mathbb{N}_0 \\ 1 & \text{otherwise.} \end{cases}$$

The extension $\bar{\varphi}$ of the function φ can indeed be seen as canonical since it holds that

$$\begin{aligned}\boldsymbol{p}^l(\bar{\varphi}) &\propto \int \bar{\varphi}(x_{\mathcal{T}_l}) p_0(x_0) \prod_{t\in\mathcal{T}_l\setminus\{T\}} K^l(x_t, x_{t+h_l}) \prod_{k=0}^{T} \ell(x_k, y_k) \mathrm{d}x_{\mathcal{T}_l} \\ &= \int \varphi(x_{0:T}) \ell(x_0, y_0) p_0(x_0) \prod_{k=1}^{T} \left[Q^l(x_{k-1}, x_k)\ell(x_k, y_k)\right] \mathrm{d}x_{0:T},\end{aligned}$$

as expected. Henceforth, $\boldsymbol{p}^l(\varphi)$ will be used has a shorthand notation for $\boldsymbol{p}^l(\bar{\varphi})$ when there is no ambiguity.

At this stage, standard Bayesian inference methods can be easily applied. For instance, if a and b are linear and constant functions respectively and if the observation equation (2.2) takes the form

$$Y_k = g_k(X_k) + V_k$$

with g_k a linear map and with V_k normally distributed, then the Kalman methodology can be used to determine the filtering and smoothing distributions. When this is not the case, the PF methodology can be used instead, the approach exposed in [9] being one of the most popular versions. The latter applies sampling and resampling mechanisms to determine the filtering distribution with an error that is uniform in time. It is however less efficient for smoothing problems [23], mostly because of the path degeneracy induced by the use of repeated resampling procedures.

The proposed second step toward the efficient computation of $\boldsymbol{p}(\varphi)$ is to use a method that enables i.i.d. samples to be drawn directly from the smoothing distribution $\boldsymbol{p}^l$ and hence avoiding path degeneracy. This has been made possible by transport methods [28, 27] which are presented in the next section.

2.3. Transport methodology. The general principle of transport methods, when applied to the considered problem, is to compute a deterministic coupling between the *base* probability distribution $\boldsymbol{\eta}^l$ of a convenient i.i.d. process on $\mathbb{R}^d$ and the *target* distribution $\boldsymbol{p}^l$, that is to compute a mapping $\boldsymbol{G}^l$ from $\mathbb{R}^{d(M_lT+1)}$ to itself that pushes forward $\boldsymbol{\eta}^l$ to $\boldsymbol{p}^l$, i.e. such that

$$\boldsymbol{p}^l(\boldsymbol{x}^l) = \boldsymbol{G}^l_{\#}\boldsymbol{\eta}^l(\boldsymbol{x}^l) \doteq \boldsymbol{\eta}^l\big((\boldsymbol{G}^l)^{-1}(\boldsymbol{x}^l)\big)\big|\det \nabla(\boldsymbol{G}^l)^{-1}(\boldsymbol{x}^l)\big|,$$

where $\nabla(\boldsymbol{G}^l)^{-1}(\boldsymbol{x}^l)$ is the gradient of the inverse transport map $(\boldsymbol{G}^l)^{-1}$ evaluated at $\boldsymbol{x}^l \in \mathbb{R}^{d(M_lT+1)}$. In this setting, the distribution $\boldsymbol{\eta}^l$ is also assumed to be on $\mathbb{R}^{d(M_lT+1)}$. The method introduced in [27] makes use of the specific structure of $\boldsymbol{p}^l$, which is induced by the Markov property of the underlying diffusion process $\boldsymbol{X}$, to divide the problem into a sequence of low-dimensional couplings. Each of these deterministic couplings, say M_t^l for some $t \in \mathcal{T}_l \setminus \{T\}$, is a mapping from $\mathbb{R}^d \times \mathbb{R}^d$ to itself which is assumed to take the form

$$M_t^l : (x_t, x_{t+h_l}) \mapsto \big(M_t^{l,1}(x_t, x_{t+h_l}), M_t^{l,2}(x_{t+h_l})\big)^{\mathrm{t}},$$

for some $M_t^{l,1} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ and $M_t^{l,2} : \mathbb{R}^d \to \mathbb{R}^d$. Under additional assumptions on $M_t^{l,1}$ and $M_t^{l,2}$ (see (2.7) below), the mapping M_t^l can be characterised by

$$(M_t^l)_{\#}\boldsymbol{\eta}^l_{t,t+h_l} = \boldsymbol{\pi}_{t,t+h_l},$$

where the probability distribution $\boldsymbol{\eta}^l_{t,t+h_l}$ on $\mathbb{R}^d \times \mathbb{R}^d$ is the marginal of $\boldsymbol{\eta}^l$ at discretization steps $(t, t+h_l)$ and where $\boldsymbol{\pi}_{t,t+h_l}$ is related to the marginal law of (X_t, X_{t+h_l}) and is characterised when $t > 0$ by

$$\boldsymbol{\pi}_{t,t+h_l}(x_t, x_{t+h_l}) \propto \begin{cases} \eta^l_t(x_t) K^l\big(M^{l,2}_{t-h_l}(x_t), x_{t+h_l}\big) \ell(x_{t+h_l}, y_{t+h_l}) & \text{if } t + h_l \in \mathbb{N} \\ \eta^l_t(x_t) K^l\big(M^{l,2}_{t-h_l}(x_t), x_{t+h_l}\big) & \text{otherwise,} \end{cases}$$

where η^l_t is the marginal of $\boldsymbol{\eta}^l$ on $\mathbb{R}^d$ at discretization time t, and by

$$\boldsymbol{\pi}_{0,h_l}(x_0, x_{h_l}) \propto \begin{cases} p_0(x_0) K^0(x_0, x_1) \ell(x_0, y_0) \ell(x_1, y_1) & \text{if } l = 0 \\ p_0(x_0) K^l(x_0, x_{h_l}) \ell(x_0, y_0) & \text{otherwise.} \end{cases}$$

Remark 2.1. The expression of $\boldsymbol{\pi}_{t,t+h_l}$ at level 0 is the one corresponding to the standard state space model presented in [27], that is

$$\boldsymbol{\pi}_{t,t+1}(x_t, x_{t+1}) \propto \eta_t(x_t) K\big(M^2_{t-1}(x_t), x_{t+1}\big) \ell(x_{t+1}, y_{t+1}), \qquad t > 0$$
$$\boldsymbol{\pi}_{0,1}(x_0, x_1) \propto p_0(x_0) K(x_0, x_1) \ell(x_0, y_0) \ell(x_1, y_1),$$

where the superscripts 0 indicating the level have been omitted.

The distribution $\boldsymbol{\eta}^l$ is a design variable which is chosen to be the normal distribution $\mathcal{N}(0, \boldsymbol{I}_{d(M_l T+1)})$ for the sake of convenience (so that $\boldsymbol{\eta}^l_{t,t+h_l} = \phi(\cdot\,; 0, \boldsymbol{I}_{2d})$ and $\eta^l \doteq \eta^l_t = \phi(\cdot\,; 0, \boldsymbol{I}_d)$ do not depend on t). The two components of the mapping M^l_t are instrumental for the proposed approach since they allow to transport samples from a convenient distribution to samples from the filtering or smoothing distributions. The filtering case is straightforward since it holds [27, Theorem 7.1] that $M^{l,2}_t$ pushes forward $\eta^l_{t+h_l}$ to the filtering distribution $p^l_{t+h_l}$. To obtain samples from the smoothing distribution, it is necessary to first embed M^l_t into the identity function on $\mathbb{R}^{d(M_l T+1)}$, which results in a function G^l_t defined as

$$G^l_t : (x_0, x_{h_l}, \ldots, x_T) \mapsto \big(x_0, \ldots, x_{t-h_l}, M^{l,1}_t(x_t, x_{t+h_l}), M^{l,2}_t(x_{t+h_l}), x_{t+2h_l}, \ldots, x_T\big)^{\mathrm{t}}.$$

It is also demonstrated in [27, Theorem 7.1] that the desired mapping $\boldsymbol{G}^l$, that is the one that pushes forward $\boldsymbol{\eta}^l$ to the smoothing distribution $\boldsymbol{p}^l$, is defined by the composition

$$\boldsymbol{G}^l = G^l_0 \circ G^l_{h_l} \circ \cdots \circ G^l_{T-h_l}. \tag{2.6}$$

Remark 2.2. It would be possible to deduce a collection $\{\tilde{G}^{l-1}_t\}_t$ of transport maps at level $l-1$ by approximating pairwise compositions of maps at level l as

$$\tilde{G}^{l-1}_t \approx G^l_t \circ G^l_{t+h_l}$$

for any $t \in \mathcal{T}_{l-1} \setminus \{T\}$. However, it is less clear in this case which distribution is approximated by this new collection of transport maps.

Although the transport maps M^l_t have been identified, their computation is not

straightforward. Assuming that the mappings $M_t^{l,1}$ and $M_t^{l,2}$ are of the form

(2.7)
$$M_t^{l,1}(x_{1:d}, x'_{1:d}) = \begin{bmatrix} M_t^{l,1,1}(x_{1:d}, x'_{1:d}) \\ \vdots \\ M_t^{l,1,d}(x_d, x'_{1:d}) \end{bmatrix} \quad \text{and} \quad M_t^{l,2}(x_{1:d}) = \begin{bmatrix} M_t^{l,2,1}(x_{1:d}) \\ \vdots \\ M_t^{l,2,d}(x_d) \end{bmatrix},$$

for any $x_{1:d}, x'_{1:d} \in \mathbb{R}^d$, i.e. loosely speaking, that $M_t^{l,1}$ and $M_t^{l,2}$ are upper triangular, it follows that M_t^l is a σ-generalised Knothe-Rosenblatt (KR) rearrangement with $\sigma = (2d, 2d-1, \dots, 1)$, that is, informally, a map whose i^{th} component depends only on the variables $x_{2d}, \dots, x_i$ and which pushes forward the i^{th} conditional of the base distribution to the corresponding conditional of the target distribution (see [27, Definition A.3] for more details). In order to find M_t^l, we first have to solve the following optimisation problem:

(2.8)
$$M^{l,*} = \operatorname*{argmin}_M -\mathbb{E}\left(\log \boldsymbol{\pi}_{t,t+h_l}(S_\sigma(M(\boldsymbol{Z}))) + \sum_{i=1}^{2d} \log \partial_i M^i(\boldsymbol{Z}) - \log \boldsymbol{\eta}^l_{t,t+h_l}(S_\sigma(\boldsymbol{Z})) \right)$$

subject to M being a monotone increasing lower triangular mapping, where the expectation is w.r.t. $\boldsymbol{Z} \sim \boldsymbol{\eta}^l_{t,t+h_l}$ and where S_σ is the linear map corresponding to the transposition matrix induced by σ. It follows that $M_t^l = S_\sigma \circ M^{l,*} \circ S_\sigma$ since it holds that $S_\sigma^{-1} = S_\sigma$ for the considered permutation σ. The above optimisation problem can be solved in different ways, e.g. by Gauss quadrature or by having recourse to Monte Carlo techniques [25, 8].

The transport map $\boldsymbol{G}^l$ enables an approximation of $\boldsymbol{p}^l(\varphi)$ to be computed by drawing N samples $\{\boldsymbol{z}_i\}_{i=1}^N$ from $\boldsymbol{\eta}^l$ and by computing the empirical average

$$\tilde{\boldsymbol{p}}^l(\varphi) \doteq \frac{1}{N} \sum_{i=1}^N \varphi\big(\boldsymbol{G}^l(\boldsymbol{z}_i)\big) \approx \boldsymbol{p}^l(\varphi).$$

The MSE corresponding to the approximation of $\boldsymbol{p}(\varphi)$ by $\tilde{\boldsymbol{p}}^l(\varphi)$ can be expressed as the sum of a variance term and a bias term as follows

$$\mathbb{E}\big((\tilde{\boldsymbol{p}}^l - \boldsymbol{p})(\varphi)^2\big) = \mathbb{E}\big((\tilde{\boldsymbol{p}}^l - \boldsymbol{p}^l)(\varphi)^2\big) + (\boldsymbol{p}^l - \boldsymbol{p})(\varphi)^2.$$

We propose to further enhance the estimation by having recourse to a multilevel strategy for which transport methods will appear to be particularly well suited.

Although the method presented in this section applies in principle to state spaces of any dimension, it is important to note that the computational cost of the corresponding algorithm can be prohibitively high even for moderate dimensions. This issue can however be mitigated by identifying some specific dependence structure between the different dimensions and by applying the same principles as the ones applied here between time steps.

3. Multilevel Monte Carlo. We now consider that the discretization (2.3) of the SDE (2.1) is performed at different discretization levels $l \in \{0, \dots, L\}$ so that $0 < h_L < \cdots < h_0 = 1$ for the considered value of h_l. This implies that the solution at the coarsest level $l = 0$ is computationally efficient but possibly inaccurate whereas the solution at the finest level L is more accurate but slower to compute. The principle of MLMC is that the respective advantages of the coarsest and finest levels can be

combined within a single estimation procedure by coupling the estimation of $\boldsymbol{p}(\varphi)$ for adjacent levels. More specifically, the first step is to notice that the smoothing distribution $\boldsymbol{p}^L$ corresponding to the discretization at level L can be expressed via a telescopic sum involving the smoothing distributions $\boldsymbol{p}^l$ at the other levels $l < L$, that is

$$\boldsymbol{p}^L(\varphi) = \sum_{l=0}^{L} (\boldsymbol{p}^l - \boldsymbol{p}^{l-1})(\varphi) \tag{3.1}$$

where $\boldsymbol{p}^{-1}$ is an arbitrary measure satisfying $\boldsymbol{p}^{-1}(\varphi) = 0$, e.g. the null measure. Equation (3.1) motivates the introduction of some i.i.d. random variables $\{\boldsymbol{X}_i^0\}_{i=1}^{N_0}$ in $\mathbb{R}^{d(T+1)}$ with law $\boldsymbol{p}^0$ and some i.i.d. random variables $\{\boldsymbol{X}_i^{l,l-1}\}_{i=1}^{N_l}$ in the space $\mathbb{R}^{d(M_l T+1)} \times \mathbb{R}^{d(M_{l-1} T+1)}$ expressed as $\boldsymbol{X}_i^{l,l-1} = (\boldsymbol{X}_i^l, \boldsymbol{X}_i^{l-})$ and such that $\boldsymbol{X}_i^l$ and $\boldsymbol{X}_i^{l-}$ have marginal laws $\boldsymbol{p}^l$ and $\boldsymbol{p}^{l-1}$ respectively, for all $l \in \{1, \dots, L\}$. This enables an approximation of $\boldsymbol{p}^L(\varphi)$ as

$$\boldsymbol{p}^L(\varphi) \approx \tilde{\boldsymbol{p}}^L(\varphi) \doteq \frac{1}{N_0} \sum_{i=1}^{N_0} \varphi(\boldsymbol{X}_i^0) + \sum_{l=1}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} \big(\varphi(\boldsymbol{X}_i^l) - \varphi(\boldsymbol{X}_i^{l-})\big). \tag{3.2}$$

This approximation of $\boldsymbol{p}^L$ is useful if the random variables $\boldsymbol{X}_{i_0}^0, \boldsymbol{X}_{i_1}^{1,0}, \dots, \boldsymbol{X}_{i_L}^{L,L-1}$ are independent of each other for all $i_0, i_1, \dots, i_L$ and if their respective components $\boldsymbol{X}_1^l$ and $\boldsymbol{X}_1^{l-}$ are as correlated as possible for all $l \in \{1, \dots, L\}$ (and hence for all random variables $\boldsymbol{X}_i^l$ and $\boldsymbol{X}_i^{l-}$ with $i \in \{1, \dots, N_l\}$ since they are i.i.d.).

In order to determine the number of samples N_l required at each level, we first express the MSE related to (3.2) as the sum of a variance term and a bias term as

$$\mathbb{E}\big((\tilde{\boldsymbol{p}}^L - \boldsymbol{p})(\varphi)^2\big) = \sum_{l=0}^{L} \mathcal{V}_l + (\boldsymbol{p}^L - \boldsymbol{p})(\varphi)^2 \tag{3.3}$$

with

$$\mathcal{V}_l = \begin{cases} \mathbb{E}\left(\left[\frac{1}{N_0} \sum_{i=1}^{N_0} \varphi(\boldsymbol{X}_i^0) - \boldsymbol{p}^0(\varphi)\right]^2\right) & \text{if } l = 0 \\ \mathbb{E}\left(\left[\frac{1}{N_l} \sum_{i=1}^{N_l} \big(\varphi(\boldsymbol{X}_i^l) - \varphi(\boldsymbol{X}_i^{l-})\big) - (\boldsymbol{p}^l - \boldsymbol{p}^{l-1})(\varphi)\right]^2\right) & \text{otherwise.} \end{cases} \tag{3.4}$$

Assuming that the bias is of order $\mathcal{O}(h_L^\alpha)$ for some integer $\alpha > 0$, it follows that a bias proportional to ϵ requires

$$L \propto -\frac{1}{\alpha} \log_2(\epsilon).$$

We also assume that the variance $\mathcal{V}_l$ at level $l > 0$ is of order $\mathcal{O}(h_l^\beta)$ and that the cost $\mathcal{C}_l$ at level l is of order $\mathcal{O}(h_l^{-\zeta})$ for some positive integers β and ζ. The number of samples N_l at level $l > 1$ can then be determined by optimising the total cost $\mathcal{C} = \sum_l \mathcal{C}_l N_l$ for a given total variance $\mathcal{V} = \sum_l \mathcal{V}_l / N_l$. This leads to

$$N_l = N_1 2^{-(\beta+\zeta)(l-1)/2}, \tag{3.4}$$

so that, to obtain a MSE of order ϵ^2, that is a bias of order ϵ and a total variance of order ϵ^2, one must take $N_0 \propto \epsilon^{-2}$ and

$$N_1 \propto \epsilon^{-2} \sum_{l=1}^{L} 2^{(\zeta-\beta)l/2}.$$

Therefore, the number of samples and the cost for a MSE of order $\mathcal{O}(\epsilon^2)$ depends on the respective values of β and ζ. For instance, if $\beta > \zeta$, then both N_1 and $\mathcal{C}$ are of order $\mathcal{O}(\epsilon^{-2})$.

3.1. Multilevel particle filter. It is assumed in this section that the interest lies in estimating the filtering distribution p_k^L at time k through the multilevel identity (3.1). Since it is generally difficult to sample directly from a reasonable candidate for a coupling of p_k^l and p_k^{l-1}, one solution is to adopt a PF strategy within the ML formulation. In order to obtain samples that are correlated between two adjacent levels, a special joint Markov transition $Q^{l,l-1}$ can be devised together with a resampling procedure that retains the correlation of the samples. This is the principle of the MLPF which is briefly discussed here. Assume that we have some collections of samples $\{x_{i,k-1}^l\}_{i=1}^{N_l}$ and $\{x_{i,k-1}^{l-}\}_{i=1}^{N_l}$ at time $k-1$ approximating p_{k-1}^l and p_{k-1}^{l-1} respectively. For all $i \in \{1,\dots,N_l\}$ and all $l \in \{1,\dots,L\}$, samples $x_{i,k}^l$ and $x_{i,k}^{l-}$ at time k are produced through the Markov transition $Q^{l,l-1}((x_{i,k-1}^l, x_{i,k-1}^{l-}),\cdot)$ as follows:

(i) Simulate (2.3) starting from the initial condition $x_0 = x_{i,k-1}^l$ over M_l time steps, denote by $x_{i,k}^l$ the obtained state of the process and by $\{u_t^l\}_{t\in\{0,h_l,\dots,1-h_l\}}$ the collection of realisations of the perturbation U_t^l drawn during the procedure.

(ii) Using the initial condition $x_0^{l-} = x_{i,k-1}^{l-}$, define $x_{i,k}^{l-}$ as the result of the deterministic recursion

$$x_{t+h_{l-1}}^{l-} = x_t^{l-} + h_{l-1}a(x_t^{l-}) + \sqrt{h_{l-1}}b(x_t^{l-})(u_t^l + u_{t+h_l}^l),$$

for any $t \in \{0, h_{l-1}, \dots, 1-h_{l-1}\}$. This recursion is meaningful since $h_{l-1} = 2h_l$ so that $u_t^l + u_{t+h_l}^l$ corresponds to the noise in the step from t to $t+h_{l-1}$ induced by $\{u_t^l\}_t$.

This procedure yields N_l pairs of correlated samples $\{(x_{i,k}^l, x_{i,k}^{l-})\}_{i=1}^{N_l}$ according to the predictive distribution at time k given observations up to time $k-1$. The information provided by the observation y_k is simply taken into account by attributing the respective weights $w_{i,k}^l$ and $w_{i,k}^{l-}$ to the samples $x_{i,k}^l$ and $x_{i,k}^{l-}$ in a similar fashion:

$$w_{i,k}^l = \frac{\ell(x_{i,k}^l, y_k)}{\sum_{j=1}^{N_l} \ell(x_{j,k}^l, y_k)} \qquad \text{and} \qquad w_{i,k}^{l-} = \frac{\ell(x_{i,k}^{l-}, y_k)}{\sum_{j=1}^{N_l} \ell(x_{j,k}^{l-}, y_k)}.$$

Following the weighting of the samples, the difference $(p_k^l - p_k^{l-1})(\varphi)$ can be estimated via

$$(p_k^l - p_k^{l-1})(\varphi) \approx \sum_{i=1}^{N_l} \Big(w_{i,k}^l \varphi\big(x_{i,k}^l\big) - w_{i,k}^{l-} \varphi\big(x_{i,k}^{l-}\big) \Big).$$

Although this approximation would behave well in general, most of the sample weights would tend to 0 if we were to apply the same procedure repeatedly in order to reach

the next observation times, resulting in a rapid increase of the empirical variance. The usual way to address this problem in the standard PF formulation is to perform resampling, that is to draw new samples from the old ones according, for instance, to the multinomial distribution induced by the weights. Applying the same approach to the MLPF would result in the loss of the correlation between the samples at adjacent levels. A *coupled* resampling is used instead as follows. For all $i \in \{1, \dots, N_l\}$ and all $l \in \{1, \dots, L\}$:

(i) With probability $\rho_k^l = \sum_{i=1}^{N_l} \min\{w_{i,k}^l, w_{i,k}^{l-}\}$ draw the index i^l according to the probability mass function (p.m.f.) $\hat{m}_k^l$ on $\{1, \dots, N_l\}$ characterised by

$$\hat{m}_k^l(j) = \frac{1}{\rho_k^l} \min\{w_{j,k}^l, w_{j,k}^{l-}\}$$

and define $i^{l-} = i^l$.

(ii) If (i) is not selected (with probability $1 - \rho_k^l$), draw the indices i^l and i^{l-} independently according to the p.m.f.s m_k^l and m_k^{l-} on $\{1, \dots, N_l\}$ characterised by

$$m_k^l(j) \propto w_{j,k}^l - \min\{w_{j,k}^l, w_{j,k}^{l-}\} \qquad \text{and} \qquad m_k^{l-}(j) \propto w_{j,k}^{l-} - \min\{w_{j,k}^l, w_{j,k}^{l-}\}.$$

(iii) Define the new pair of samples $(\tilde{x}_{i,k}^l, \tilde{x}_{i,k}^{l-})$ as $(x_{i^l,k}^l, x_{i^{l-},k}^{l-})$.

Although the *coupled* resampling addresses the problem of reducing the empirical variance without completely losing the correlation between samples at adjacent levels, it nevertheless has a negative impact of the ML rate. Indeed, as demonstrated in [20], one needs $\beta > 2\zeta$ to obtain a cost of order $\mathcal{O}(\epsilon^{-2})$ for a MSE of order $\mathcal{O}(\epsilon^2)$. In the case where $\beta = 2\zeta$, e.g. for Euler's scheme ($\zeta = 1$) with $\beta = 2$, the cost is of order $\mathcal{O}(\epsilon^{-2} \log(\epsilon)^2)$.

Also, even if the MLPF can handle smoothing on a short time window, i.e. it can successfully approximate the distribution of $\{X_{t'}\}_{t' \in \{t-s, t-s+1, \dots, t\}}$ given $y_0, \dots, y_t$ for small values of $s \in \mathbb{N}$, the error in the approximation of the full smoothing distribution would increase in time because of the path degeneracy effect. Indeed, resampling tends to multiply the samples of higher weights so that, after a certain number of time steps, all samples will be descendants of the same earlier sample.

3.2. Multilevel transport. In order to avoid the path degeneracy inherent to any PF approach and to regain the ML rate lost through the coupled resampling of the MLPF, we propose to compute samples from the distributions $\boldsymbol{p}^l$ via the transport maps $\boldsymbol{G}^l$ characterised by $\boldsymbol{p}^l = \boldsymbol{G}_\#^l \boldsymbol{\eta}^l$ with $\boldsymbol{\eta}^l = \phi(\cdot; 0, \boldsymbol{I}_{d(M_l T+1)})$ for all $l \in \{0, \dots, L\}$. The specific procedure is described as follows. For all $i \in \{1, \dots, N_l\}$:

(i) draw a sample $\boldsymbol{z}_i^l = (z_{i,0}^l, z_{i,1}^l, \dots, z_{i,M_l T}^l)$ from $\boldsymbol{\eta}^l$

(ii) map $\boldsymbol{z}_i^l$ through $\boldsymbol{G}^l$ to obtain a sample $\boldsymbol{x}_i^l = \boldsymbol{G}^l(\boldsymbol{z}_i^l)$ from $\boldsymbol{p}^l$

(iii) define a *thinned* sample $\boldsymbol{z}_i^{l-} = (z_{i,0}^l, z_{i,2}^l, \dots, z_{i,M_l T}^l)$

(iv) map $\boldsymbol{z}_i^{l-}$ through $\boldsymbol{G}^{l-1}$ to obtain a sample $\boldsymbol{x}_i^{l-} = \boldsymbol{G}^{l-1}(\boldsymbol{z}_i^{l-})$ from $\boldsymbol{p}^{l-1}$

This simple procedure yields two collections $\{\boldsymbol{x}_i^l\}_i$ and $\{\boldsymbol{x}_i^{l-}\}_i$ of samples drawn from a joint distribution that obviously has marginals $\boldsymbol{p}^l$ and $\boldsymbol{p}^{l-1}$ and that correlates adjacent levels as desired. As a motivation for this coupling, note that it is optimal in terms of squared Wasserstein distance with the Euclidean metric in the case where $d = 1$ and assuming that the transport maps can be computed exactly. The efficiency of the approach comes from the fact that the transport maps $\boldsymbol{G}^l$ have to be computed once only. Given the computation of the maps, it is relatively fast to obtain the samples.

Although there is, strictly speaking, no path degeneracy in the considered approach, there might be some accumulation of error through time induced by the composition of transport maps defining $\boldsymbol{G}^l$ as in (2.6). This accumulation of error will however be seen to be milder than the one experienced by the PF in section 4.

It is assumed that the procedure underlying the computation of the transport maps is deterministic, so that there is no undesired correlations between samples from $\boldsymbol{X}^{l,l-1}$ and $\boldsymbol{X}^{l',l'-1}$ when $l \neq l'$. Further neglecting the numerical error in the computed transport maps, it follows that the expression (3.3) of the MSE holds for the considered approach.

Before proceeding to a numerical study, the legitimacy of the proposed approach is verified for the linear-Gaussian case. Consider the SDE (2.1) in dimension $d = 1$ and with $p_0 = \delta_{x_0}$ (so that the observation at time $t = 0$ has no impact). The corresponding filtering distribution at time $k \in \mathbb{N}$ and at level $l \in \{0, \dots, L\}$ simplifies to

$$p_k^l(x_k) \propto \int \prod_{n=1}^{k} \big[Q^l(x_{n-1}, x_n)\ell(x_n, y_n)\big] \mathrm{d}x_{1:k-1}$$

for any $x_k \in \mathbb{R}^d$. Denote $\hat{G}_k^l \doteq M_{k-h_l}^{l,2}$ the transport map from the base distribution $\eta^l = \phi(\cdot\,; 0, 1)$ to p_k^l, i.e. such that $(\hat{G}_k^l)_\# \eta^l = p_k^l$. If F_{η^l} and $F_{l,k}$ denote the cumulative distribution functions (c.d.f.) of η^l and p_k^l respectively, then it holds that $\hat{G}_k^l = F_{l,k}^{-1} \circ F_{\eta^l}$, where F^{-1} is the generalised inverse

$$F^{-1}(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}, \qquad \forall u \in [0, 1].$$

Considering i.i.d. random variables $Z_i \sim \eta^l$ for $i \in \{1, \dots, N_l\}$, the objective is to determine the order of

$$\mathcal{V}_{l,k} = \mathrm{Var}\left(\frac{1}{N_l}\sum_{i=1}^{N_l}\Big(\varphi\big(\hat{G}_k^l(Z_i)\big) - \varphi\big(\hat{G}_k^{l-1}(Z_i)\big)\Big)\right)$$

w.r.t. h_l for any function φ that is at the intersection of the set $\mathcal{B}_b(\mathbb{R})$ of bounded measurable functions and of the set $\mathrm{Lip}(\mathbb{R})$ of Lipschitz functions. Since the Z_i's are i.i.d. and by definition of $\hat{G}_k^l$, it holds that

$$\begin{aligned}
\mathcal{V}_{l,k} &= \frac{1}{N_l}\mathrm{Var}\Big(\varphi\big(\hat{G}_k^l(Z)\big) - \varphi\big(\hat{G}_k^{l-1}(Z)\big)\Big) \\
&= \frac{1}{N_l}\mathrm{Var}\Big(\varphi\big(F_{l,k}^{-1}(U)\big) - \varphi\big(F_{l-1,k}^{-1}(U)\big)\Big) \\
&\leq \frac{c}{N_l}\mathbb{E}\Big(\big[F_{l,k}^{-1}(U) - F_{l-1,k}^{-1}(U)\big]^2\Big)
\end{aligned}$$

for some $c > 0$, with $Z \sim \eta^l$ and $U \sim \mathcal{U}([0, 1])$, where the inequality comes from the fact that $\varphi \in \mathrm{Lip}(\mathbb{R})$. The linear case is addressed in the following theorem as a proof of concept.

THEOREM 3.1. *Let $\boldsymbol{X}$ a 1-dimensional diffusion process with linear drift and constant diffusion coefficient observed at all integer times through a linear-Gaussian likelihood $\ell(x, \cdot) = \phi(\cdot\,; x, \tau^2)$ for some $\tau > 0$, then the variance $\mathcal{V}_{l,k}$ obtained at level l for Euler's method with discretization $h_l = 2^{-l}$ and with the transport-based approach*

satisfies

$$\mathcal{V}_{l,k} = \mathcal{O}(h_l^2)$$

for any $k \in \{1, \dots, T\}$.

Proof. The objective is to compute the order of

$$F_{l,k}^{-1}(u) - F_{l-1,k}^{-1}(u) = \hat{\mu}_{l,k} - \hat{\mu}_{l-1,k} + \sqrt{2}\,\mathrm{erf}^{-1}(2u-1)(\hat{\sigma}_{l,k} - \hat{\sigma}_{l-1,k})$$

w.r.t. h_l, where erf^{-1} is the inverse error function and where the updated mean $\hat{\mu}_{l,k}$ and standard deviation $\hat{\sigma}_{l,k}$ at level l and at time k can be found through the Kalman filter to be

$$\hat{\mu}_{l,k} = \mu_{l,k} + \frac{\sigma_{l,k}^2(y - \mu_{l,k})}{\tau^2 + \sigma_{l,k}^2} \qquad \text{and} \qquad \hat{\sigma}_{l,k}^2 = \frac{\tau^2 \sigma_{l,k}^2}{\tau^2 + \sigma_{l,k}^2}$$

with $\mu_{l,k}$ and $\sigma_{l,k}$ the predicted mean and standard deviation expressed as

$$\mu_{l,k} = (1 + h_l a)^{M_l} \hat{\mu}_{l,k-1} \qquad \text{and} \qquad \sigma_{l,k}^2 = (1 + h_l a)^{2M_l} \hat{\sigma}_{l,k-1}^2 + h_l b^2 \sum_{i=0}^{M_l - 1} (1 + h_l a)^{2i}.$$

First, the predicted mean $\mu_{l,k}$ and standard deviation $\sigma_{l,k}$ have to be developed to the second order. The main term appearing in the expressions of $\mu_{l,k}$ is

$$(1 + h_l a)^{M_l} = \sum_{n=0}^{M_l} \frac{a^n}{n!} \prod_{i=0}^{n-1} \big[h_l(M_l - i)\big] = \sum_{n=0}^{M_l} \frac{a^n}{n!} + \frac{h_l}{2} \sum_{n=2}^{M_l} \frac{a^n}{(n-2)!} + \mathcal{O}(h_l^2),$$

For the sake of compactness we define

$$A_m = \sum_{n=0}^{m} \frac{a^n}{n!} \qquad \text{and} \qquad B_m = \sum_{n=2}^{m} \frac{a^n}{(n-2)!}.$$

Assuming that

$$\hat{\mu}_{l,k-1} = c_{k-1} + r_{k-1,l} h_l + \mathcal{O}(h_l^2) \tag{3.7a}$$

$$\hat{\sigma}_{l,k-1} = c'_{k-1} + r'_{k-1,l} h_l + \mathcal{O}(h_l^2) \tag{3.7b}$$

where c_{k-1} and c'_{k-1} do not depend on l, and where $r_{k-1,l}$ and $r'_{k-1,l}$ are of order $\mathcal{O}(1)$ w.r.t. h_l, it follows that

$$\begin{aligned} \mu_{l,k} &= \hat{\mu}_{l,k-1}\Big(A_{M_l} + \frac{h_l}{2} B_{M_l}\Big) + \mathcal{O}(h_l^2) \\ &= c_{k-1} A_{M_l} + r_{k-1,l} h_l A_{M_l} + h_l \frac{c_{k-1}}{2} B_{M_l} + \mathcal{O}(h_l^2). \end{aligned}$$

Recalling that $M_l = 2^l$ and noticing that

$$A_{M_l} = e^a - \sum_{n \geq M_l + 1} \frac{a^n}{n!} = e^a + o(h_l)$$

with $o(h_l)$ referring to terms that are negligible in front of h_l, $\mu_{k,l}$ can be seen to be of the same form as $\hat{\mu}_{k,l}$, that is

$$\mu_{l,k} = c_{k-1} e^a + r_{k-1,l} h_l e^a + h_l \frac{c_{k-1}}{2} B_{M_l} + \mathcal{O}(h_l^2).$$

The same type of expansion can be used for the first term in the variance $\sigma_{l,k}^2$ as follows

$$\sigma_{l,k}^2 = c_{k-1}'^2 e^a + 2c_{k-1}' r_{k-1,l}' h_l e^a + h_l \frac{c_{k-1}'^2}{2} B_{2^{l+1}} + b^2 h_l \sum_{i=0}^{M_l-1} (1+h_l a)^{2i} + \mathcal{O}(h_l^2).$$

The second term has however a slightly different form and must be studied on its own:

$$h_l \sum_{i=0}^{M_l-1} (1+h_l a)^{2i} = \sum_{n=0}^{2M_l-2} \left(h_l^{n+1} a^n \sum_{i=\lceil n/2 \rceil}^{M_l-1} \binom{2i}{n} \right) \tag{3.9}$$

where it appears that

$$h_l^{n+1} a^n \sum_{i=\lceil n/2 \rceil}^{M_l-1} \binom{2i}{n} \leq h_l^{n+1} a^n \sum_{i=1}^{M_l} \frac{(2i)^n}{n!} = \frac{(2a)^n}{(n+1)!}$$

where the r.h.s. tends exponentially fast to 0 when $n \to \infty$. It follows that (3.9) is of the form $s + o(h_l)$ where s does not depend on l, so that

$$\sigma_{l,k}^2 = c_{k-1}'^2 e^a + 2c_{k-1}' r_{k-1,l}' h_l e^a + h_l \frac{c_{k-1}'^2}{2} B_{2^{l+1}} + sb^2 + \mathcal{O}(h_l^2),$$

from which the expansion of the standard deviation $\sigma_{l,k}$ can be expressed as

$$\sigma_{l,k} = \sqrt{C_l} + \frac{h_l}{2\sqrt{C_l}} \left(2c_{k-1}' r_{k-1,l}' h_l e^a + \frac{c_{k-1}'^2}{2} B_{2^{l+1}} \right) + \mathcal{O}(h_l^2)$$

where $C_l = e^a c_{k-1}'^2 + sb^2$ is the term of order $\mathcal{O}(1)$ in $\sigma_{l,k}^2$. We conclude that

$$\mu_{l,k} - \mu_{l-1,k} = h_l \big(r_{k-1,l} A_{2^l} - 2 r_{k-1,l-1} A_{2^{l-1}} \big) + h_l \frac{c_{k-1}}{2} \big(B_{2^l} - 2 B_{2^{l-1}} \big) + \mathcal{O}(h_l^2) = \mathcal{O}(h_l).$$

Similarly, it holds that $\sigma_{l,k} - \sigma_{l-1,k} = \mathcal{O}(h_l)$. Proceeding to the updated terms, it holds that

$$\begin{aligned} \sigma_{l,k}^2 (y_k - \mu_{l,k}) &= (e^a c_{k-1}'^2 + sb^2)(y_k - c_{k-1} e^a) + \mathcal{O}(h_l) \\ \tau^2 + \sigma_{l,k}^2 &= \tau^2 + (e^a c_{k-1}'^2 + sb^2) + \mathcal{O}(h_l), \end{aligned}$$

so that

$$\begin{aligned} \hat{\mu}_{l,k} &= \mu_{l,k} + \frac{\sigma_{l,k}^2 (y_k - \mu_{l,k})}{\tau^2 + \sigma_{l,k}^2} = c_{k-1} e^a + \frac{(e^a c_{k-1}'^2 + sb^2)(y_k - c_{k-1} e^a)}{\tau^2 + (e^a c_{k-1}'^2 + sb^2)} + \mathcal{O}(h_l) \\ \hat{\sigma}_{l,k} &= \frac{\tau^2 \sigma_{l,k}^2}{\tau^2 + \sigma_{l,k}^2} = \frac{\tau^2 (e^a c_{k-1}'^2 + sb^2)}{\tau^2 + (e^a c_{k-1}'^2 + sb^2)} + \mathcal{O}(h_l). \end{aligned}$$

If follows from reasoning by induction that $\hat{\mu}_{l,k}$ and $\hat{\sigma}_{l,k}$ have the form assumed in (3.7) for all $k \in \{0, \ldots, T\}$, the result being obvious for $k = 0$. Combining the different results it can be easily verified that

$$\hat{\mu}_{l,k} - \hat{\mu}_{l-1,k} = \mathcal{O}(h_l) \qquad \text{and} \qquad \hat{\sigma}_{l,k} - \hat{\sigma}_{l-1,k} = \mathcal{O}(h_l),$$

which yields $\mathcal{V}_{l,k} = \mathcal{O}(h_l^2)$ as desired. This concludes the proof of the theorem. □

4. Numerical study. In this section, the effectiveness of the proposed method is shown in simulations for different SDE models. Numerical verifications of some of the considered assumptions are also provided. The scenarios considered for simulation are the same as for the MLPF in [20], so that results can be compared.

4.1. Linear Gaussian. The first simulation study is performed on the linear-Gaussian case with $a = -0.1$, $b = 1$ and with a likelihood $\ell(x, \cdot) = \phi(\cdot\,; x, \tau^2)$ with $\tau = 0.25$ which corresponds to an observation process of the form

$$Y_k \,|\, X_k \sim \mathcal{N}(0, \tau^2). \tag{4.1}$$

The initial distribution is $p_0 = \phi(\cdot\,; 0, \sigma)$ with $\sigma = 1$ and the final time is $T = 4$. A realisation of the state and observation processes are shown in Figure 4.1 together with the mean and some percentiles corresponding to samples drawn from the smoothing distribution. The involved transport maps[1], say T, are assumed to be triangular maps which i^{th} component $T^{(i)}$ takes the form

$$T^{(i)}(x_1, \ldots, x_i) = a_i(x_1, \ldots, x_{i-1}) + \int_0^{x_i} b_i(x_1, \ldots, x_{i-1}, t)^2 \mathrm{d}t$$

where a_i and b_i are real-valued functions defined on $\mathbb{R}^{i-1}$ and $\mathbb{R}^i$ respectively. For any $j \leq i-1$, it is assumed that the functions $x_j \mapsto a_i(x_1, \ldots, x_{i-1})$ and $x_j \mapsto b_i(x_1, \ldots, x_{i-1}, t)$ are Hermite Probabilists' functions extended with constant and linear components whereas the function $t \mapsto b_i(x_1, \ldots, x_{i-1}, t)$ is assumed to be a Hermite Probabilists' function extended with a constant component only. Then, the functions a_i and b_i, when expressed as functions from $\mathbb{R}^{i-1}$ and $\mathbb{R}^i$ respectively, take the form

$$a_i(x_1, \ldots, x_{i-1}) = \sum_{k=1}^{2d(o_{\mathrm{m}}+1)} c_k \Phi_k(x_1, \ldots, x_{i-1})$$

$$b_i(x_1, \ldots, x_{i-1}, t) = \sum_{k=1}^{2do_{\mathrm{m}}} c'_k \Psi_k(x_1, \ldots, x_{i-1}, t)$$

with o_{m} the map order, with $\{c_k\}_{k\geq 1}$ and $\{c'_k\}_{k\geq 1}$ some collections of real coefficients and with Φ_k and Ψ_k basis functions based on the above mentioned Hermite Probabilists' functions. In the simulations, the case $o_{\mathrm{m}} = 4$ is considered.

The integration in (2.8) is performed using a Gauss quadrature of order 10 in each dimension. The optimisation relies on the Newton-CG algorithm (Newton algorithm using the conjugate-gradient method for each step) with a tolerance of 10^{-4}.

MLMC rates. The behaviour of the numerical scheme for different levels is displayed in Figure 4.2a, where $\mathrm{Var}(\varphi(\boldsymbol{X}^l) - \varphi(\boldsymbol{X}^{l-1}))$ is considered with $\varphi(x_{0:T}) = x_T$ and where the cost is the computational time required to obtain one sample at a given level l. This result confirms the applicability of multilevel techniques by showing that $\mathcal{V}_l = \mathcal{O}(h_l^2)$ and $\mathcal{C}_l = \mathcal{O}(h_l^{-1})$, that is $\beta = 2$ and $\zeta = 1$.

One important point is that the time spent to obtain samples at a high level is small when compared to the time required to compute the underlying transport map. For instance, it takes about 25s to calculate the transport map at level 5 while a sample

[1] The solver used for the determination of the transport maps is the one provided at http://transportmaps.mit.edu/docs/index.html

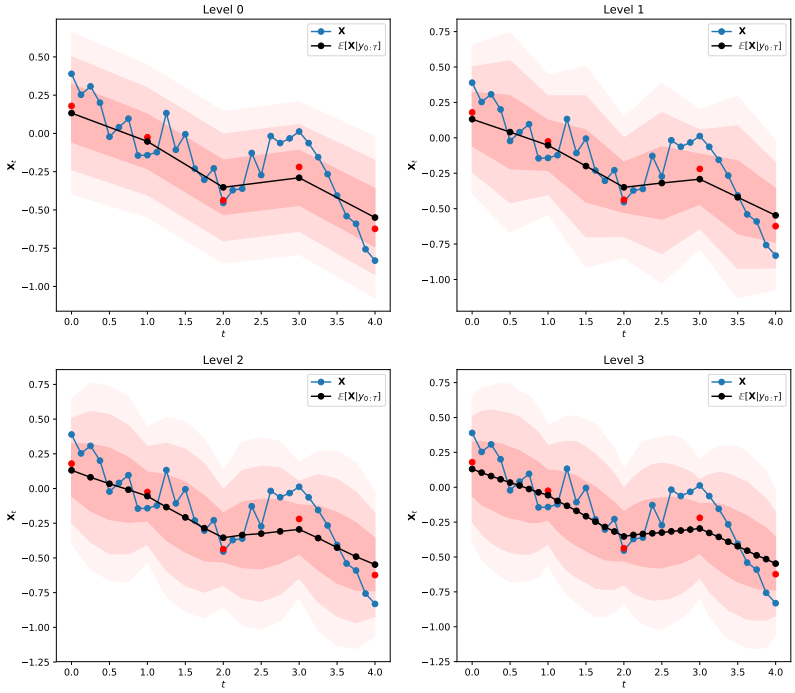
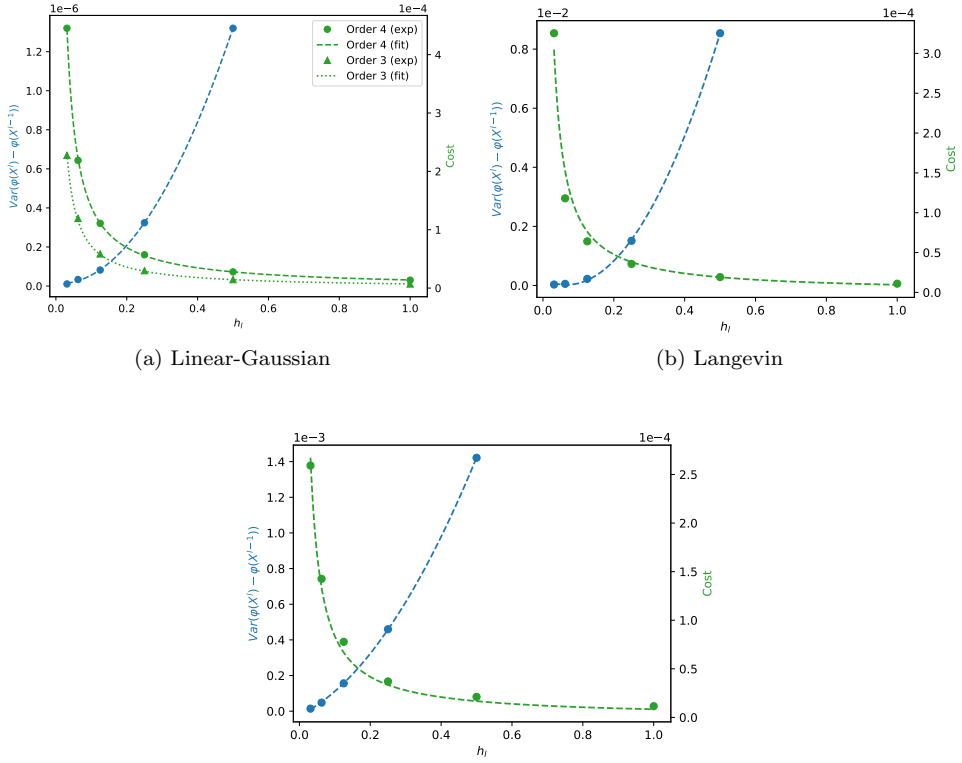


FIG. 4.1. *Mean and percentiles of samples generated according to the target distribution of the linear-Gaussian SDE at four consecutive levels (blue line: state of the process; red dots: observations; black line: samples mean; red areas:* 1-99, 5-95 *and* 20-80 *percentiles).*

is obtained in 0.00025s, so that a 100,000 samples can be drawn in the time spent to compute the map. It is therefore necessary to verify that the gain obtained with the multilevel approach is not compensated by the additional time spent computing more transport maps (one for each level).

Multilevel vs computation at the highest level. The objective with the multilevel approach is to reduce the computational cost to reach a given error when compared to computations at the highest level only. This aspect is verified in Figure 4.3a where the multilevel approach appears to outperform the one based on samples at the highest level. The above-mentioned fact that calculation of the transport maps might be time-consuming is shown to be compensated by the efficiency of the multilevel approach within a reasonable time interval. This is in spite of the fact that the multilevel approach nearly doubles the number of maps to be computed. In particular, in the considered linear-Gaussian scenario, the average computational cost for the calculation of the maps in the multilevel and highest-level approach is respectively 10.76s and 6.15s.

More specifically, Figure 4.3 is obtained by first computing all the required transport maps and then by generating samples by batches of 1000. The multilevel estimate is obtained by sweeping the different levels sequentially until the predetermined number N_l of samples has been computed at level l. The number N_0 of samples at level 0 is fixed to $2^{13} \times 1000$ for all the considered SDEs, that is 2^{13} batches of 1000 samples. The number of samples at level 1 is determined by the ratio between the variance at levels 0 and 1 and the number of samples for the subsequent levels are computed through (3.4).



(a) Linear-Gaussian

(b) Langevin

(c) Non-linear diffusion

FIG. 4.2. *Variance of* $\varphi(\boldsymbol{X}^l) - \varphi(\boldsymbol{X}^{l-1})$ *with* $\varphi(x_{0:T}) = x_T$ *and cost as a function of* h_l *(Blue dashed line: poly. fit of order* 2*; green dashed line: least-square fitting of the form* a/h_l*). The experimental (exp) cost for two map-approximation orders are indicated in the linear-Gaussian case together with their corresponding least-square fittings (fit).*

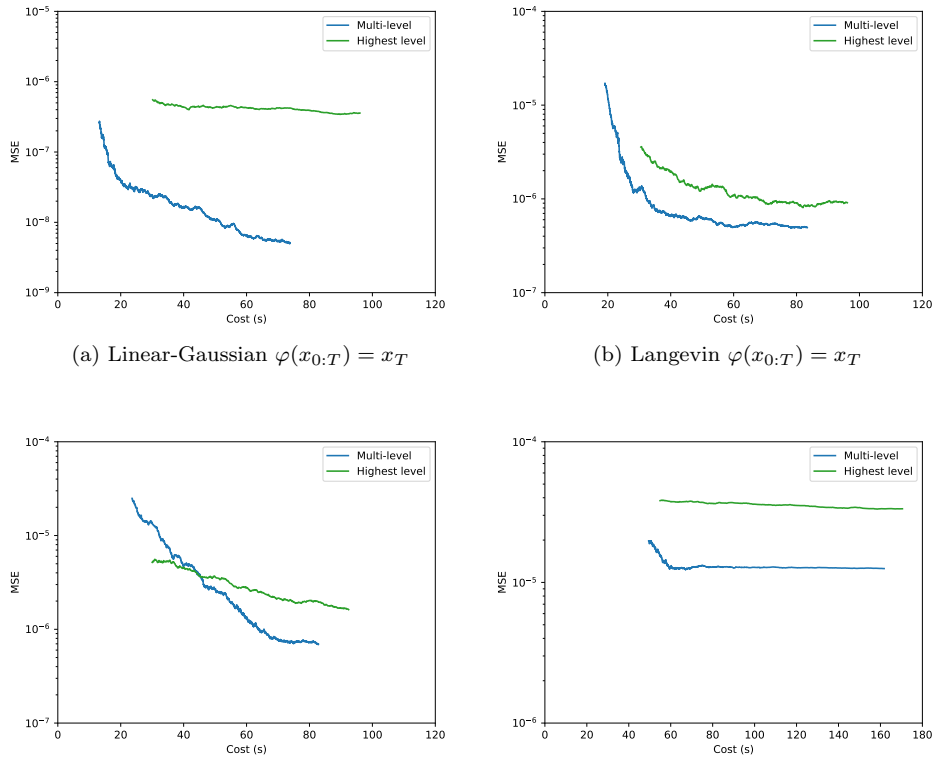
4.2. Langevin SDE. We now consider a Langevin SDE of the form

$$\mathrm{d}X_t = \frac{1}{2}\nabla \log \mathcal{S}_\nu(X_t)\mathrm{d}t + b\,\mathrm{d}W_t, \qquad t \in [0, T]$$

where $\mathcal{S}_\nu$ is the Student's t distribution with $\nu = 10$ degrees of freedom and with $b = 1$. The observations are generated according to

$$Y_k \,|\, X_k \sim \mathcal{N}\big(0, \tau^2 \exp(X_k)\big) \tag{4.2}$$

with $\tau = 1$. The initial distribution is the same as in the previous example. A realisation of the considered Langevin SDE is shown in Figure 4.4 together with mean and percentiles of samples obtained using transport maps. It appears clearly on this figure that the observation process characterised by (4.2) is less informative than the one modelled by (4.1). Figure 4.2b shows that the considered Langevin SDE also displays a variance of order $\mathcal{O}(h_l^2)$, although the actual values are much higher than in the linear-Gaussian case, which might be due to both the nature of the SDE and the quality of the approximation of the transport maps. A comparison of the computational efficiency of the multilevel approach is given in Figure 4.3b where the



(a) Linear-Gaussian $\varphi(x_{0:T}) = x_T$

(b) Langevin $\varphi(x_{0:T}) = x_T$

(c) Langevin $\varphi(x_{0:T}) = \sum_{t=0}^{T} e^{-\kappa(T-t)} x_t$

(d) Non-linear diffusion $\varphi(x_{0:T}) = x_T$

FIG. 4.3. *MSE vs. cost for the multilevel approach compared with computations at the highest level $L = 4$ (semi-log scale, averaged over 50 Monte Carlo simulations). The first* 200 *iterations are not displayed.*

proposed method is seen to outperform the approach based on computations at the highest level. The time needed to initialise the latter, i.e. the time to compute the transport map at level $L = 4$ and to perform the first 200 iterations, is however slightly less affected than with the multilevel approach. Figure 4.3c shows the performance of the proposed approach with a different functional, that is

$$\varphi(x_{0:T}) = \sum_{t=0}^{T} \exp(-\kappa(T-t)) x_t,$$

which gives the sum of the states at the observations weighted by a forgetting factor κ, with $\kappa = 2$ in the simulations. In this case, the tolerance of the optimisation is also adapted to the level as follows: the tolerance at level l is 10^{-l-1}. This helps retaining the benefits of the multi-level approach in this more challenging smoothing problem.

4.3. Nonlinear diffusion. We now consider a SDE with a nonlinear diffusion term:

$$\mathrm{d}X_t = \theta(\mu - X_t)\mathrm{d}t + \frac{\varsigma}{\sqrt{1 + X_t^2}}\mathrm{d}W_t, \qquad t \in [0, T]$$

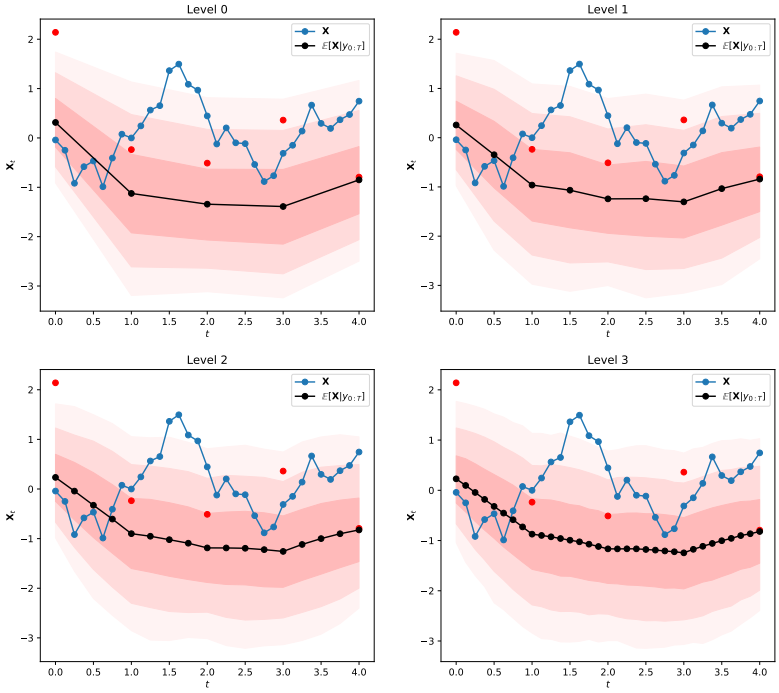


FIG. 4.4. *Mean and percentiles of samples generated according to the target distribution of the Langevin SDE at four consecutive levels (blue line: state of the process; red dots: observations; black line: samples mean; red areas:* 1-99, 5-95 *and* 20-80 *percentiles).*

with $\theta = 1$, $\mu = 1$ and $\varsigma = 1$ and with a time step of 0.5 between observation times, so that the final time is $T = 2$. The linear-Gaussian observation model (4.1) is considered with $\tau = 1$. The initial distribution is the same as in the previous examples. A realisation of the considered SDE is displayed in Figure 4.5 together with mean and percentiles of samples obtained using transport maps. Figure 4.2c shows that the same rates as in the previous cases apply although the contribution of the quadratic term in the variance is smaller than before. It appears in Figure 4.3d that the time spent computing the transport maps has largely increased for both approaches when compared to the linear-Gaussian and Langevin SDEs. This might be due to the challenging nature of the problem which induces a slower convergence on the involved optimisation methods. However, the proposed method still displays a significant gain in performance, although the first 200 iterations just gave it enough time to compensate for the computational overhead caused by the calculation of the maps at all level.

5. Conclusion. An algorithm for the determination of expectations with respect to laws of partially-observed SDEs has been proposed. The observations are received at discrete times and depend only on the state at the time they occurred, hence enabling a standard state space modelling to be used. The proposed method relies on three principles: (i) the discretization of the considered SDE, for instance with Euler's method, (ii) the expression of the smoothing distribution at a given level as a telescopic sum involving coarser discretizations and (iii) the generation of pairs of samples correlated across adjacent levels via the application of different transport maps to samples from a common base distribution. As opposed to MLPF, the pro-

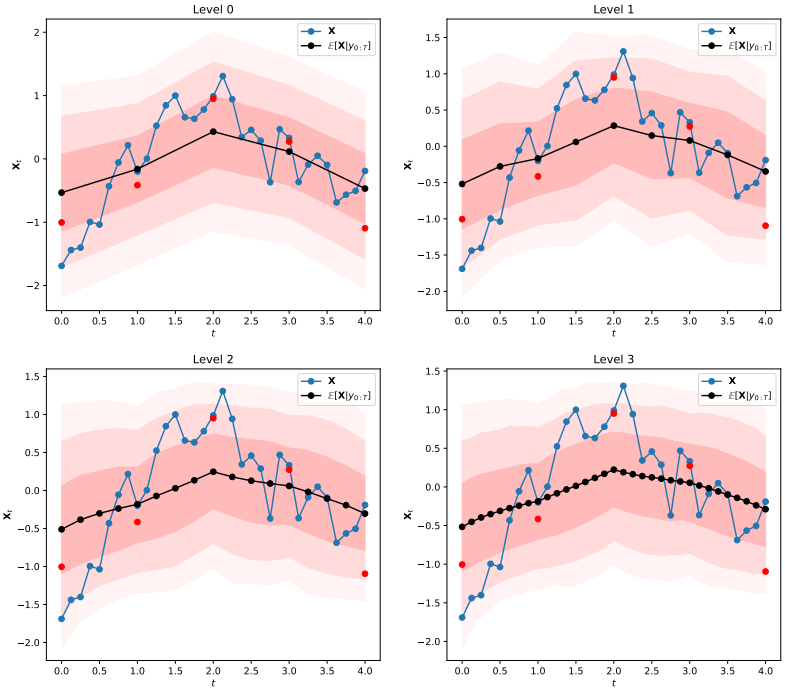


FIG. 4.5. *Mean and percentiles of samples generated according to the target distribution of the SDE with nonlinear diffusion at four consecutive levels (blue line: state of the process; red dots: observations; black line: samples mean; red areas:* 1-99*,* 5-95 *and* 20-80 *percentiles).*

posed approach retains the "ideal" MLMC rates, since, in particular, it does not require resampling techniques to be used. In addition to a numerical verification of its performance, the proposed method was shown to have the desired behaviour in the linear-Gaussian case. Future works include the theoretical verification of the rates that are observed in practice for more diverse types of SDEs, as well as the study of the optimal parametrisation of the transport maps as a function of the discretization level.

Acknowledgements. The authors would like to thank the Associate Editor as well as the referees for their detailed comments and suggestions for the manuscript. All authors were supported by Singapore Ministry of Education AcRF tier 1 grant R-155-000-182-114. AJ was also supported under CRG4 Award Ref:2584. AJ is affiliated with the Risk Management Institute, OR and analytics cluster and the Center for Quantitative Finance at NUS.

REFERENCES

[1] Y. AIT-SAHALIA, *Closed-form likelihood expansions for multivariate diffusions*, The Annals of Statistics, 36 (2008), pp. 906–937.
[2] A. BESKOS, O. PAPASPILIOPOULOS, G. O. ROBERTS, AND P. FEARNHEAD, *Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion)*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68 (2006), pp. 333–382.
[3] A. BESKOS AND G. O. ROBERTS, *Exact simulation of diffusions*, The Annals of Applied Probability, 15 (2005), pp. 2422–2444.

[4] O. Cappe, E. Moulines, and R. T, *Inference in Hidden Markov Models*, Springer, 2005.
[5] D. Crisan and A. Bain, *Fundamentals of Stochastic Filtering*, Springer, 2008.
[6] D. Crisan, P. Del Moral, J. Houssineau, and A. Jasra, *Unbiased multi-index Monte Carlo*, Stoch. Anal. Appl., 36 (2018), pp. 257–273.
[7] D. Crisan and B. Rozovskii, *The Oxford handbook of nonlinear filtering*, OUP, 2011.
[8] P. J. Davis and P. Rabinowitz, *Methods of numerical integration*, Courier Corporation, 2007.
[9] A. Doucet and A. Johansen, *A tutorial on particle filtering and smoothing: Fifteen years later*, In Handbook of Nonlinear Filtering, OUP, (2011).
[10] T. A. El Moselhy and Y. M. Marzouk, *Bayesian inference with optimal maps*, Journal of Computational Physics, 231 (2012), pp. 7815–7850.
[11] P. Fearnhead, O. Papaspiliopoulos, and G. O. Roberts, *Particle filters for partially observed diffusions*, J. R. Statist. Soc. Ser. B, 70 (2008), pp. 755–777.
[12] M. B. Giles, *Multilevel Monte Carlo path simulation*, Operations Research, 56 (2008), pp. 607–617.
[13] M. B. Giles, *Multilevel Monte Carlo methods*, Acta Numerica, 24 (2015), pp. 259–328.
[14] M. B. Giles, D. J. Higham, and X. Mao, *Analysing multi-level Monte Carlo for options with non-globally Lipschitz payoff*, Finance and Stochastics, 13 (2009), pp. 403–413.
[15] M. B. Giles and L. Szpruch, *Antithetic multilevel Monte Carlo estimation for multi-dimensional SDEs without lévy area simulation*, The Annals of Applied Probability, 24 (2014), pp. 1585–1620.
[16] A. Gregory, C. Cotter, and S. Reich, *Multilevel ensemble transform particle filtering*, SIAM J. Sci. Comp., 38 (2016), pp. A1317–A1338.
[17] A.-L. Haji-Ali, F. Nobile, and R. Tempone, *Multi-index Monte Carlo: when sparsity meets sampling*, Numerische Mathematik, 132 (2016), pp. 767–806.
[18] S. Heinrich, *Multilevel Monte Carlo methods*, In Large-Scale Scientific Computing Springer, (2001).
[19] J. Heng, A. Doucet, and Y. Pokern, *Gibbs flow for approximate transport with applications to Bayesian computation*, arXiv preprint arXiv:1509.08787, (2015).
[20] A. Jasra, K. Kamatani, K. J. Law, and Y. Zhou, *Multilevel particle filter*, SIAM J. Numer. Anal., 55 (2017), pp. 3068–3096.
[21] A. Jasra, K. Kamatani, P. Osei, and Y. Zhou, *Multilevel particle filter: normalizing constant estimation*, Stat. Comp., 28 (2018), pp. 47–60.
[22] A. Jasra, K. J. H. Law, and C. Suciu, *Advanced Multilevel Monte Carlo*, arXiv preprint arXiv:1704.07272v1, (2017).
[23] N. Kantas, A. Doucet, S. S. Singh, J. M. Macijowski, and N. Chopin, *On particle methods for parameter estimation in general state-space models*, Statist. Sci., 20 (2015), pp. 328–351.
[24] M. Parno, T. Moselhy, and Y. Marzouk, *A multiscale strategy for Bayesian inference using transport maps*, SIAM/ASA Journal on Uncertainty Quantification, 4 (2016), pp. 1160–1190.
[25] C. P. Robert, *Monte Carlo methods*, Wiley Online Library, 2004.
[26] D. Sen, A. Thiery, and A. Jasra, *On coupling particle filters*, Stat. Comp., 28 (2018), pp. 461–475.
[27] A. Spantini, D. Bigoni, and Y. Marzouk, *Inference via low-dimensional couplings*, arXiv preprint arXiv:1703.06131, (2017).
[28] C. Villani, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.